Research on Prediction of Decision Tree Algorithm on Different Types of Stocks

Shipei Du, Xiao Li and Dongjie Yang
Guangdong University of Science and Technology, Dongguan, China

Keywords: Decision Tree, Stock Prices, Prediction.

Abstract: The stock market is an important part of the financial market and is closely connected to a country's market growth and economic patterns. Because of how much the stock market changes and the non-linear nature of it, accurately guessing what will happen in the stock market is really hard. In this article, we suggest a model for predicting the stock market using a decision tree algorithm. We will use historical trading data from multiple A-shares for our research. We use artificial intelligence and the decision tree algorithm to study the financial industry and make predictions about stock prices. Our research found that using the decision tree algorithm to predict stocks gave us good results. This is helpful information for guiding both big institutions and individuals in making stock investments.

1 INTRODUCTION

In recent years, with the widespread application of artificial intelligence technology in the financial field, the decision tree algorithm, as a simple and effective machine learning algorithm, has become one of the important tools in the financial field. In terms of stock market forecasting, the decision tree algorithm has high accuracy and reliability, which can help investors better understand market trends and stock price changes.

Decision tree algorithm is a machine learning algorithm based on model selection and decision tree construction, which solves complex problems by decomposing them into a series of simple problems and building decision trees 0. In stock market forecasting, decision tree algorithms can use historical transaction data for stock price forecasting and analysis 0. By building a decision tree model, we can comprehensively analyze various factors that affect stock prices, such as market conditions, company fundamentals, and technical indicators, and obtain stock price forecast results.

Forecasting the final price of stocks can greatly enhance the profits made from investing in stocks and help shareholders make better decisions about their investments 0. Stock price is the core variable of the stock market, which is affected by many factors, such as economic factors, policy factors, company fundamentals, market sentiment, etc. By predicting the closing price of stocks, investors can better grasp the trend of stock prices, thereby better grasping investment opportunities and reducing risks 0. Therefore, predicting the closing price of stocks has an important guiding role in the investment decisions of shareholders.

The research object of this research is the historical transaction data of multiple A-share stocks. We have selected a number of representative stocks, including large-cap blue-chip stocks, mid- and small-cap stocks, and emerging industry stocks. Through the analysis of the historical transaction data of these stocks, we found that the decision tree algorithm has a better effect in stock market forecasting 0. By constructing a decision tree model, we can comprehensively analyze various factors that affect the stock price, and obtain the forecast results of the stock price 0.

This research applies the decision tree algorithm to the analysis of A-share historical transaction data, and uses the decision tree model to predict future stock prices, aiming to provide valuable insights and guidance for financial institutions and individual investors.

Our research contributes to the growing literature on applied AI techniques in finance 0. We believe that the decision tree algorithm is a powerful tool that can improve the accuracy and efficiency of stock market forecasting. Our findings have important value to financial institutions and

Du, S., Li, X. and Yang, D.
Research on Prediction of Decision Tree Algorithm on Different Types of Stocks.
DOI: 10.5220/0012277000003807

In *Proceedings of the 2nd International Seminar on Artificial Intelligence, Networking and Information Technology (ANIT 2023)*, pages 178-181
ISBN: 978-989-758-677-4

individual investors seeking to make informed investment decisions.

2 DATASET AND METHOD

In this research, the dataset used comes from the Shanghai Composite Index data of the first half of 2022 in China, and 6 stocks are selected as research objects, including 2 blue-chip stocks, 2 Small and medium cap stocks, and 2 emerging industry stocks. They are China Telecom (stock code: 601728), Ping An Insurance (Group)(stock code: 601318), Beijing Tongrentang (stock code: 600085), China Zheshang Bank (stock code: 601916), Jiangxi Hongcheng Environment (stock code: 600461), and Humanwell Healthcare (Group) (stock code: 600079). For ease of description, they are coded as BC-1, BC-2, SM-1, SM-2, EI-1, and EI-2, respectively.

For the training data set D, use the least squares regression tree generation algorithm to output the decision tree $f(x)$. In the input space where the training data set is located, recursively divide each area into two sub-areas and determine the output value on each sub-area, and construct a binary decision tree. Select the optimal segmentation variable j and segmentation point s, and solve:

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2\right] \quad (1)$$

Iterates over variable j, scans split point s for fixed split variable j, partitions the region with the selected pair (j,s) and determines the corresponding output value:

$$R_1(j,s)=\{x|\ x^{(j)}\le s\},\quad R_2(j,s)=\{x|\ x^{(j)}>s\} \quad (2)$$

$$\hat{c}_m=\frac{1}{N_m}\sum_{x_i\in R_m(j,s)}y_i\quad,\quad x\in R_m,\quad m=1,2 \quad (3)$$

Continue to call formulas (1), (2), (3) for both subregions until the stop condition is met. Divide the input space into M regions R_1, R_2, ... R_M and generate the decision tree:

$$f(x)=\sum_{m=1}^{M}\hat{c}_m I(x\in R_m) \quad (4)$$

In this research, the data set was analyzed and processed for data such as date, opening, high point, low point, closing, and trading volume. Decision tree machine learning algorithms were used to predict stock trends. This dataset was divided into a training set and a test set, with the total number of datasets being 702, with a training set-to-test set ratio of 85% to 15%. For a single stock, the total number of samples was 117, with a training set of 100 samples and a test set of 17 samples. The maximum depth of the decision tree was set to 100 layers, and the minimum number of samples at a leaf node was set to 2. Regression forecasting models were utilized, and data analysis techniques were employed to accurately explain the pattern of data movement and create visual representations of the analysis outcomes. Additionally, we used different evaluation indicators to perform performance analysis on our prediction model.

Table 1: The real and predicted values of BC-1, BC-2, SM-1, and the deviation.

Sample	BC-1 real	BC-1 prediction	BC-1 deviation	BC-2 real	BC-2 prediction	BC-2 deviation	SM-1 real	SM-1 prediction	SM-1 deviation
1	4	4.01	0.25%	48.45	49.03	1.19%	46.74	46.75	0.02%
2	3.85	3.85	0.13%	43.66	43.05	1.40%	35.49	35.44	0.15%
3	4.13	4.13	0.00%	51.46	51.74	0.54%	42.84	44.32	3.46%
4	3.75	3.74	0.40%	46.13	45.99	0.30%	51	52.68	3.28%
5	4.03	4.02	0.25%	49.54	49.03	1.04%	45.94	45.84	0.21%
6	4	3.98	0.42%	48.47	48.50	0.06%	42.92	41.29	3.81%
7	3.91	3.86	1.21%	44	44.27	0.61%	45.68	45.98	0.66%
8	4.14	4.10	1.09%	50.92	51.05	0.26%	46.42	45.84	1.24%
9	4.19	4.16	0.66%	54.16	53.78	0.71%	43.63	43.46	0.39%
10	3.75	3.75	0.00%	44.14	44.65	1.15%	49.77	45.84	7.89%
11	4.3	4.35	1.16%	52.07	51.25	1.58%	48.4	49.42	2.11%
12	4.08	4.03	1.23%	48.25	47.44	1.69%	43.36	43.37	0.02%
13	4.41	4.35	1.36%	53.04	54.17	2.14%	45.43	45.98	1.22%
14	4.3	4.30	0.08%	51.3	51.25	0.11%	51.45	49.00	4.76%
15	4.31	4.35	0.93%	55.59	53.78	3.26%	42.97	44.97	4.65%
16	3.91	3.92	0.34%	44.15	44.14	0.02%	46.58	45.84	1.58%
17	3.85	3.86	0.26%	43.29	43.76	1.09%	40.99	41.23	0.59%

3 EXPERIMENT AND ANALYSIS

A type of algorithm based on a tree structure is called a "decision tree," which can be used to make a series of decisions and ultimately reach the best conclusion 0. This is a supervised learning algorithm that is suitable for both classification and regression problems. In classification problems, the decision tree divides the data set into different classes; in regression problems, the decision tree predicts a continuous value. In general, decision trees use

information entropy or information gain as a measure of how to partition data. Although the advantage of decision trees is that they are easy to understand and can visually represent the decision-making process, overfitting is a problem that needs to be paid attention to 0. The quality and quantity of training data are crucial for decision trees.

In this study, the information is separated into two groups: one for training and the other for testing. There are 702 pieces of information for 6 stocks in total. For one stock, there are 117 pieces of information. The training set contains 85% of the data, and the test set contains 15% of the data. There are 100 examples in the training group and 17 examples in the testing group. The results of the prediction can be seen in Table 1 and Table 2.

Table 2: The real and predicted values of SM-2, EI-1, EI-2, and the deviation.

Sample	SM-2 real	SM-2 prediction	SM-2 deviation	EI-1 real	EI-1 prediction	EI-1 deviation	EI-2 real	EI-2 prediction	EI-2 deviation
1	3.32	3.34	0.70%	7.59	7.49	1.32%	17.25	17.11	0.83%
2	3.25	3.21	1.38%	7.77	7.68	1.19%	14.2	13.73	3.35%
3	3.44	3.43	0.22%	7.98	8.08	1.25%	18.86	19.13	1.41%
4	3.36	3.38	0.67%	8.04	8.03	0.16%	17.29	17.58	1.65%
5	3.42	3.38	1.10%	7.75	7.94	2.45%	16.94	17.30	2.10%
6	3.33	3.34	0.40%	7.57	7.55	0.31%	15.25	14.41	5.51%
7	3.25	3.26	0.31%	8.05	8.15	1.28%	16.16	16.42	1.62%
8	3.47	3.50	0.86%	8.21	8.23	0.18%	18.43	18.37	0.34%
9	3.52	3.50	0.57%	8.27	8.22	0.60%	19.19	18.99	1.04%
10	3.28	3.31	0.76%	8.07	8.01	0.78%	16.15	15.59	3.47%
11	3.52	3.50	0.57%	8.5	8.57	0.76%	23.59	24.21	2.64%
12	3.28	3.29	0.41%	7.21	7.17	0.52%	17.23	17.06	0.97%
13	3.55	3.55	0.00%	8.2	8.18	0.21%	20.18	19.53	3.21%
14	3.5	3.50	0.00%	8.58	8.57	0.17%	24.65	24.21	1.77%
15	3.57	3.56	0.28%	8.23	8.22	0.12%	19.01	20.11	5.80%
16	3.28	3.21	2.29%	8	8.05	0.63%	16.58	16.49	0.56%
17	3.28	3.25	0.84%	8.08	8.11	0.31%	16.02	16.52	3.12%

It can be seen from the above table that most of the deviations between the predicted values of the six stocks and the real values are less than 3%. The curves of the predicted values and the real values of the six stocks are shown in Figure 1.

MSA(Mean Squared Absolute Error), RMSE(Root Mean Squared Error), MAE(Mean Absolute Error), and R^2(R-squared) are all commonly used regression error metrics or scoring indicators. MSA is calculated as the square root of the mean of the squared differences between the

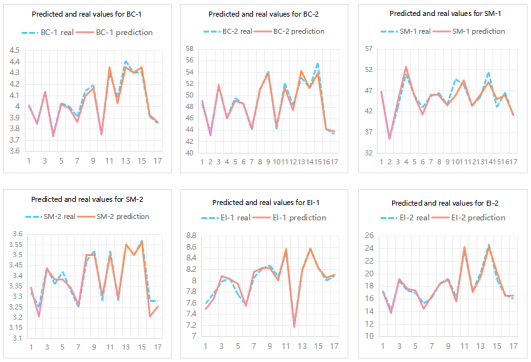

Figure 1: The curves of the predicted value and the real value of the six stocks.

predicted and true values. RMSE is calculated as the square root of the average of the squared differences between the predicted and true values. MAE is calculated as the average of the absolute values of the differences between the predicted value and the true value. R^2 is calculated as the ratio of the square of the regression coefficient to the sum of the squares of the residual. The value of R^2 is between 0 and 1, and the closer to 1, the better the fitting effect of the model. For the above six stocks, the forecasting performance indicators are shown in Table 3.

Table 3: The performance of the decision tree model.

Stock	MSE	RMSE	MAE	R^2	Average deviation
BC-1	0.001	0.035	0.026	0.968	0.57%
BC-2	0.379	0.616	0.481	0.972	1.01%
SM-1	1.707	1.306	0.927	0.907	2.12%
SM-2	0.001	0.031	0.024	0.915	0.67%
EI-1	0.008	0.091	0.062	0.918	0.72%
EI-2	0.349	0.591	0.428	0.944	2.32%

As shown in Table 3, the R^2 values of the decision tree model on all 6 stocks are close to 1, and the prediction performance of the model is good, while the Average deviation of SM-1 and EI-2 is obviously different from that of the other 4 stocks, which may be It is related to the industry background of these two stocks. These two stocks are Beijing Tongrentang and Humanwell Healthcare, both of which are in the pharmaceutical industry. This also gives us an inspiration. For different industries, machine learning algorithms may obtain different prediction performance. This is also one of our future research directions.

4 CONCLUSION

This study uses a type of technology called decision tree in machine learning to anticipate the behavior of stocks. It focuses on well-established stocks, smaller to medium-sized stocks, and stocks from new industries. It uses past data from the first six months of 2022 to guess what the closing stock prices will be. The results of the experiment show that the decision tree model is better at predicting how the stock will change in the future, and it does a good job of making accurate predictions.

The way the price of stocks goes up and down is the main thing about the stock market. Many things can affect it, like the economy, policies, how companies are doing, and how people feel about the market. These things will affect stock prices in different ways, causing the prices to go up and down. So, figuring out how to use these factors to make better predictions about stocks is something that can be looked into more in the future.

ACKNOWLEDGMENTS

This research was funded by the Social Science Project of Guangdong University of Science and Technology(GKY-2022KYYBW-6), Humanities and Social Science Youth Program of Guangdong Provincial Department of Education (2018WQNCX206).

REFERENCES

Rath S, Gupta B K, Nayak A K. Stock Market Prediction Using Supervised Machine Learning Algorithm[C]//*Advances in Distributed Computing and Machine Learning: Proceedings of ICADCML 2021*. Springer Singapore, 2022: 374-381.

Zhang C, Sjarif N N A, Ibrahim R B. Decision Fusion for Stock Market Prediction: A Systematic Review[J]. *IEEE Access*, 2022.

Sakhare N N, Shaik I S, Saha S. Prediction of stock market movement via technical analysis of stock data stored on blockchain using novel History Bits based machine learning algorithm[J]. *IET Software*, 2022.

Lee C S, Cheang P Y S, Moslehpour M. Predictive analytics in business analytics: decision tree[J]. *Advances in Decision Sciences*, 2022, 26(1): 1-29.

Zi R, Jun Y, Yicheng Y, et al. Stock price prediction based on optimized random forest model[C]//*2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML)*. IEEE, 2022: 777-783.

Illa P K, Parvathala B, Sharma A K. Stock price prediction methodology using random forest algorithm and support vector machine[J]. *Materials Today: Proceedings*, 2022, 56: 1776-1782.

Kebonye N M, Agyeman P C, Biney J K M. Optimized modelling of countrywide soil organic carbon levels via an interpretable decision tree[J]. *Smart Agricultural Technology*, 2023, 3: 100106.

Kurani A, Doshi P, Vakharia A, et al. A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting[J]. *Annals of Data Science*, 2023, 10(1): 183-208.

Behera J, Pasayat A K, Behera H, et al. Prediction based mean-value-at-risk portfolio optimization using machine learning regression algorithms for multi-national stock markets[J]. *Engineering Applications of Artificial Intelligence*, 2023, 120: 105843.

Shaikh B, Iyer A, Koneti M, et al. Stock Price Prediction with Sentimental Analysis[C]//*2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*. IEEE, 2022: 1632-1638.

Yun K K, Yoon S W, Won D. Interpretable stock price forecasting model using genetic algorithm-machine learning regressions and best feature subset selection[J]. *Expert Systems with Applications*, 2023, 213: 118803.

Srinu Vasarao P, Chakkaravarthy M. Time series analysis using random forest for predicting stock variances efficiency[M]//*Intelligent Systems and Sustainable Computing: Proceedings of ICISSC 2021*. Singapore: Springer Nature Singapore, 2022: 59-67.

Deng S, Xiao C, Zhu Y, et al. Dynamic forecasting of the Shanghai Stock Exchange index movement using multiple types of investor sentiment[J]. *Applied Soft Computing*, 2022, 125: 109132.

Lombardo G, Pellegrino M, Adosoglou G, et al. Machine Learning for Bankruptcy Prediction in the American Stock Market: Dataset and Benchmarks[J]. *Future Internet*, 2022, 14(8): 244.